Synthetische koppeling in de praktijk

A.H. Paape[1] en J.S. Cramer[2]

Samenvatting: Synthetische koppeling van microdata is een oplossing om informatie over variabelen uit verschillende steekproeven bijeen te brengen. Tot op heden is er meer over de theorie geschreven dan dat er daadwerkelijke koppelingen - met een practisch doel - werden uitgevoerd. Wij brengen twee synthetische koppelingen tot stand ten behoeve van een specifiek onderzoeksdoel. Hiervoor gebruiken wij twee verwante technieken die hun oorsprong vinden in de besliskunde. Eén voor op een mainframe-computer, de ander voor op een PC.

1 Stichting voor Economisch Onderzoek der Universiteit van Amsterdam (SEO)
2 Faculteit der Economische Wetenschappen en Econometrie, Universiteit van Amsterdam

Correspondentie-adres:
Stichting voor Economisch Onderzoek der Universiteit van Amsterdam
Roetersstraat 18
1018 WB Amsterdam
telefoon: 020-6242412

1 Inleiding

Sociale of economische gegevens op individueel niveau worden op allerlei gebied en met uiteenlopende doelstellingen door enquêtes, tellingen of metingen verzameld. Soms liggen specifieke onderzoeksvragen ten grondslag aan de verzameling van dergelijke micro-gegevens, maar zij dienen vaker voor andere doeleinden. Zo worden de budgetenquêtes onder huishoudens van het CBS gebruikt bij de samenstelling van prijsindices, en ook wel bij de raming van het verbruik van bepaalde goederen. Onderzoekers zijn vaak aangewezen op secundair gebruik van dergelijke bestanden, en de gegevens die voor een bepaalde onderzoeksvraag nodig zijn komen soms niet alle in één en hetzelfde bestand voor. Het ligt voor de hand hier in te voorzien door de informatie uit verschillende bronnen op het individuele niveau aan elkaar te *koppelen*. Bij volledige bestanden kan men overwegen de gegevens van identieke eenheden (personen, huishoudens, personenauto's etc.) uit verschillende bestanden samen te voegen, zoals bij de controle van de Motorrijtuigenbelasting bij het kentekenregister van auto's. Bij het gebruik van steekproeven is deze *directe* (of *exacte*) koppeling echter nauwelijks mogelijk, nog afgezien van de vraag of hij niet in strijd is met de bescherming van de privacy omdat de geënquêteerde op grond van de gecombineerde gegevens gemakkelijker valt te identificeren. Men kan echter ook individuele eenheden uit twee bestanden koppelen op grond van de overeenkomst tussen gemeenschappelijke kenmerken. Dit is de *synthetische* koppeling van microdata, waarbij alleen bij uitzonderlijk toeval (en ongewild) informatie van een en dezelfde onderzoekseenheid uit twee enquêtes bijeen wordt gebracht. De nieuw gevormde eenheid is als regel een kunstmatige schepping, op louter statistische gronden uit de twee enquêtes geconstrueerd.

Het is de vraag of men aan deze koppeling wat heeft. De koppeling berust altijd op een beperkt aantal gezamenlijke variabelen die in beide bestanden voorkomen, de *koppelvariabelen*, en er wordt geen rekening gehouden met de waarden van de *unieke variabelen* die telkens in één bestand voorkomen. Conditioneel op de koppelvariabelen zijn deze unieke variabelen derhalve in het geconstrueerde bestand statistisch onafhankelijk; iedere samenhang die zij vertonen berust geheel op de gezamenlijke samenhang met de koppelvariabelen. Dit is de Conditional Independence Assumption, die impliciet aan iedere synthetische koppeling ten grondslag ligt (Sims, 1972). Voor *causale analyses* is de methode dus niet geschikt, want ze voegt niets toe aan de samenhang van de unieke variabelen met de koppelvariabelen, ieder binnen hun eigen bestand. Voor de *beschrijving van een populatie* naar meerdere kenmerken kan de methode echter goede diensten verlenen. Men verkrijgt een 'benchmark' bestand dat representatief is in die zin dat alle marginale verdelingen kloppen. Zo kan tenminste worden geconcludeerd uit de omvangrijke

literatuur over het onderwerp.[1] Daarbij wekt het wel enig wantrouwen dat er veel meer over koppeling is en wordt geschreven dan er aan koppeling wordt gedaan. In Nederland vonden wij slechts één voorbeeld van een daadwerkelijke koppeling met een praktisch doel, namelijk in de combinatie van gegevens uit kijk- en luisteronderzoeken (die volgens een oud taboe altijd gescheiden worden ondernomen) door Bronner (1989).[2]

2 Het koppelingsprobleem

Technisch luidt het vraagstuk dat bij synthetische koppeling moet worden opgelost als volgt: men beschikt over twee steekproeven A en B uit dezelfde populatie met respectievelijk m en n waarnemingen. De m individuele waarnemingen i uit A hebben een steekproefgewicht of ophoogfactor a_i en bevatten de variabelen $\mathbf{x}_i$ en $\mathbf{y}_i$; de n waarnemingen j uit B hebben een steekproefgewicht b_j en bevatten de variabelen $\mathbf{x}_j$ en $\mathbf{z}_j$. De **x** zijn gemeenschappelijke variabelen, de **y** en **z** uniek. De gewichten of ophoogfactoren a_i en b_j dienen om de samenstelling van de steekproeven in het geval van stratificatie in overeenstemming te brengen met de populatie. Zij geven aan voor hoeveel eenheden in de populatie één waarneming in de steekproef model staat. Bij geslaagde aselecte trekking zijn de a_i en b_j dus constanten voor alle i respectievelijk j, en als de twee steekproeven even groot zijn dan zijn ze nog gelijk ook.

Bij de synthetische koppeling onderscheidt men A en B vaak in een basisbestand of *recipiëntbestand* (R) en een *donorbestand* (D) waarmee R wordt verrijkt. Men neemt een waarneming uit R en zoekt daar een waarneming uit D bij die er op grond van de waarden van de gemeenschappelijke **x** (zo veel mogelijk) mee overeen komt. De waarden van **z** worden aan (**x**,**y**) toegevoegd. Zo ontstaat een synthetische waarneming met $\mathbf{x}_i$ (of $\mathbf{x}_j$), $\mathbf{y}_i$ én $\mathbf{z}_j$, en gewicht w_{ij}. Als criterium voor de overeenkomst in de gemeenschappelijke variabelen **x** dient een afstandsfunctie d_{ij} van $\mathbf{x}_i$ en $\mathbf{x}_j$.

In de literatuur zijn vele koppelingstechnieken te vinden. Eén onderscheid is of de koppeling zonder of met restricties plaats vindt. Bij het koppelen zonder restricties geldt dat een waarneming uit D desgewenst meer malen aan verschillende waarnemingen uit R wordt toegevoegd: koppeling met terugleggen. Bij koppeling met restricties wordt iedere waarneming uit D niet

---

1 Zie Cramer en Paape (1990)

2 Strikt genomen is ook dit geen synthetische koppeling: Bronner beschikt over een uitvoerige enquête uit 1986 en een beperkte enquête uit 1987, en vult de laatste aan door extrapolatie van de eerste.

meer dan eens gebruikt: koppeling zonder terugleggen. Dit laatste wordt in de literatuur algemeen aanbevolen, en wij zullen ons er toe beperken.

Ook in dit geval doen zich nog problemen voor. Het kan voorkomen dat er voor een gegeven waarneming uit R geen goed gelijkende waarneming in D te vinden is, hetzij omdat die eenvoudig ontbreekt, hetzij omdat hij al eerder is gebruikt. Om dit gevaar te beperken verdient het aanbeveling de grootste steekproef als donor te gebruiken. Ook dan blijft de vraag in welke volgorde men voor de waarnemingen uit R een partner uit D zoekt.

Dit bezwaar vervalt als men de twee bestanden symmetrisch behandelt, zonder de bestanden A en B te onderscheiden in R en D. Men beschouwt nu *combinaties* van paren records uit de beide bestanden, met als nevenvoorwaarde dat het gekoppelde bestand dezelfde marginale verdelingen heeft als de oorspronkelijke (gewogen) steekproeven. Het koppelingsvraagstuk wordt dan een lineair programmeringsprobleem, met name een **transportprobleem**. Kadane (1978) en vooral ook Barr en Turner (1981a, 1981b) waren de eersten die het koppelingsvraagstuk op deze wijze benaderden. Deze wijze van koppelen brengt de optimale koppeling ("optimal constrained match") tot stand, gegeven een bepaalde afstandsfunctie d_{ij}.

Als transportprobleem ziet het koppelingsprobleem er als volgt uit:

$$\text{minimaliseer} \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} w_{ij} \qquad (1)$$

onder de voorwaarden:

$$\sum_{j=1}^{n} w_{ij} = a_i \qquad (i = 1, \ldots, m) \qquad (2)$$

$$\sum_{i=1}^{m} w_{ij} = b_j \qquad (j = 1, \ldots, n) \qquad (3)$$

$$w_{ij} \geq 0 \qquad \text{voor alle i en j} \qquad (4)$$

en:

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \qquad (5)$$

Vergelijking (5) geeft aan dat (we aannemen dat) de twee steekproeven uit dezelfde populatie zijn getrokken. Na sommatie leveren de steekproefgewichten / ophoogfactoren voor beide bestanden, die niet even groot hoeven te zijn, dezelfde marginale verdelingen. Dit vergt in de praktijk vaak enige aanpassing. De voorwaarden (2) en (3) geven aan dat de gewichten w_{ij} in het gekoppelde bestand zo bepaald moeten worden dat elke waarneming, gesommeerd over de waarnemingsparen waarin het voorkomt, even zwaar meetelt als in het bestand waaruit het vandaan komt. Deze gewichten w_{ij} zijn niet-negatief, maar wel vaak nul - namelijk als het paren niet wordt gekoppeld.

Het koppelingsprobleem aldus gedefinieerd is met behulp van algoritmen uit de besliskunde op te lossen, misschien met de omvang van de gebruikte databestanden (nog) niet op de personal computer, maar toch zeker op een mainframe. Wij hebben echter ook gekeken naar een alternatief, dat wel op een PC kan worden opgelost.

3 Een alternatief voor de PC

Een alternatief is het koppelingsprobleem te herschrijven als een **Lineair Toewijzingsprobleem** (Linear Assignment Problem of LAP). Natuurlijk levert dit geen optimale oplossing. Een groot bezwaar is dat de marginale verdeling van de unieke variabelen in het gekoppelde bestand meestal niet gelijk is aan de verdeling in het bestand van herkomst. Er gaat aldus informatie verloren.

In het LAP wordt wel verschil gemaakt tussen een recipiënt en een donor. Wij kiezen het kleinste van twee bestanden, met m waarnemingen, als R en het grootste, met n waarnemingen, als D. Aan iedere waarneming i uit R wordt nu de best gelijkende waarneming j uit D toegevoegd. Het gekoppelde bestand bevat nu dus net zo veel waarnemingen als R, en een deel van de waarnemingen uit D blijft ongebruikt. Men vindt de marginale verdeling uit het donorbestand doorgaans dan ook niet terug in het gekoppelde bestand.

Het lineaire toewijzingsprobleem luidt als volgt:

$$\text{minimaliseer} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} w_{ij} \qquad (6)$$

onder de voorwaarden:

$$\sum_{j=1}^{n} w_{ij} = a_i \qquad (i = 1, \ldots, m) \qquad (7)$$

$$\sum_{i=1}^{n} w_{ij} = b_j \qquad (j = 1, \ldots, n) \qquad (8)$$

$$\text{met } w_{ij} = \begin{cases} 1 \text{ als record j van D wordt gekoppeld aan record i van R} \\ 0 \text{ anders} \end{cases} \qquad (9)$$

Dit komt overeen met een transportprobleem voor twee aselecte steekproeven van dezelfde omvang uit dezelfde populatie, zoals men inziet als in het transportprobleem m en n aan elkaar gelijk zijn en alle a_i en b_j gelijk aan 1. Doorgaans zijn de twee steekproeven echter niet precies even groot. Om het koppelingsprobleem toch als LAP op te lossen vullen wij - bij n > m - de "afstandenmatrix" d_{ij} aan met rijen nullen voor i=m+1,....,n. Nu is d_{ij}=0 voor i=m+1,....,n en j=1,....,n, en het maakt niet uit welke w_{ij}'s één zijn voor i tussen m+1 en n.

4 Een voorbeeld: de koppeling van LVO en PAP

In de praktijk levert het oplossen van het transportprobleem moeilijkheden op. In het onderstaande hebben we bijvoorbeeld een donorbestand van circa 2.500 waarnemingen en een recipiëntbestand van circa 1.000 waarnemingen. Dit leidt bij het oplossen van het koppelingsprobleem als transportprobleem tot het optimeren van een functie van (nxm)=2,5 miljoen variabelen onder (n+m)=3.500 restricties. Dit kan een PC (nog) niet aan.

Koppelen met het lineair toewijzingsprobleem komt neer op het vinden van de optimale permutatie van (m van) de n waarnemingen van de donor over de m waarnemingen van de recipiënt. Hiervoor zijn n! / (n-m)! mogelijke oplossingen - met m=1.000 en n=2.500 is dit gelijk aan $3{,}77 \times 10^{4.843}$. Toch zijn de berekeningen eenvoudiger dan voor het transportprobleem.

We hebben beide methoden toegepast op bestanden uit twee enquêtes op het gebied van verkeer en vervoer, namelijk het *Longitudinaal Verplaatsingsonderzoek (LVO)* en het *Personenauto Panelonderzoek (PAP)*.
Het Longitudinaal Verplaatsingsonderzoek (LVO) werd gehouden in opdracht van het Directoraat-Generaal voor het Verkeer en het Projectbureau Integrale Verkeers- en Vervoerstudies van het Ministerie van Verkeer en Waterstaat. Vanaf 1984 en 1989 zijn tweemaal per jaar gegevens verzameld bij een panel van huishoudens. Het bestand bevat veel gegevens over de huishoudens: algemene karakteristieken (grootte en samenstelling, postcode, totaal huishoudinkomen, etc.) en aantal specifieke gegevens op het gebied van het verkeer (zoals het bezit van vervoermiddelen). Bovendien zijn van alle personen boven de 12 jaar in het huishouden veel individuele gegevens opgenomen, waaronder autobezit, gebruik van abonnementen of kaartsoorten voor het Openbaar Vervoer, en ook (zeer minitieuze) gegevens over verplaatsingen - maar geen jaarkilometrage van de auto. Wij gebruiken het bestand van maart 1989.

Het Personenauto Panelonderzoek (PAP) wordt gehouden door het CBS. Aan een steekproef van autobezitters wordt gevraagd vier maal achtereen, telkens een maand later, de tellerstand van de auto te noteren. Dit levert een zeer betrouwbare schatting op van het gereden aantallen kilometers van de Nederlandse automobilisten. Het bestand bevat een groot aantal gegevens over de auto, maar slechts weinig over de persoon of het huishouden waartoe de auto behoort.

Bij enig nadenken zal men inzien dat een **algemene** koppeling van twee bestanden niet mogelijk is, omdat het van het doel van de koppeling afhangt welke gemeenschappelijke variabelen als koppelvariabelen **x** worden gebruikt. Hiervoor moet men immers variabelen kiezen die sterk met de unieke variabelen samenhangen, liefst zó sterk dat er geen partiële correlatie tussen **y** en **z** overblijft - in het gekoppelde bestand zijn zij namelijk altijd conditioneel onafhankelijk.
Van een grootscheepse koppeling ter constructie van één, voor vele doeleinden bruikbaar microbestand kan dus geen sprake zijn.

In het onderhavige voorbeeld zijn de unieke variabelen het jaarlijks kilometrage (uit het PAP) en het opleidingsniveau van de automobilist (uit het LVO). De vraag luidt dus of hoger opgeleide automobilisten meer of minder kilometers rijden dan lager opgeleiden, **zonder** correctie voor andere kenmerken.

Tot dusver hebben wij stilzwijgend aangenomen dat alle bestanden betrekking hebben op dezelfde elementen. Wil men steekproeven koppelen die in dit opzicht verschillen dan moet men dit eerst door transformatie bereiken. Nu is het LVO een bestand van huishoudens, het PAP een

bestand van (particuliere) personenauto's. Het is echter mogelijk het LVO-bestand te transformeren tot een bestand van personenauto's met de gegevens van de hoofdgebruiker van deze auto's en van het huishouden waartoe de hoofdgebruikers behoort.
De koppelvariabelen zijn gekozen door binnen beide bestanden de correlaties te berekenen van de gemeenschappelijke variabelen met de unieke variabele. Aan de hand van deze cijfers kiezen wij 7 koppelvariabelen, die alle sterk of redelijk sterk met beide unieke variabelen samenhangen:

- Bouwjaar auto
- Aankoopprijs auto
- Leeftijd hoofdgebruiker auto
- Beroep / bezigheid hoofdgebruiker auto
- Aantal werkuren per week hoofdgebruiker auto
- Reistijd woonadres | werkadres hoofdgebruiker auto
- Netto-inkomen hoofdgebruiker auto.

Als maatstaf voor de afstand tussen $\mathbf{x}_i$ en $\mathbf{x}_j$ nemen wij de (gewogen) absolute afstandsfunctie:

$$d_{ij} = \sum_{k=1}^{r} G_k \, | \, x_{jk} - x_{ik} \, | \qquad (10)$$

met r het aantal koppelvariabelen, hier 7. De koppelvariabelen zijn vooraf gestandaardiseerd (gemiddelde nul, standaardafwijking één) om de afstanden van de verschillende variabelen vergelijkbaar te maken. De gewichten G_k geven de relatieve belangrijkheid van de verschillende koppelvariabelen aan; deze is hier bepaald als het gemiddelde van de correlaties met de twee unieke variabelen.
Aangezien de twee bestanden verschillende jaren betreffen hebben wij nog enige correcties op de koppelvariabelen toegepast. De autoprijzen en inkomens zijn met prijsindices tot één jaar herleid, en in plaats van het bouwjaar is de leeftijd van de auto genomen.

Na de beschreven transformatie van het LVO-bestand en het weglaten van waarnemingen die voor de koppelvariabelen "onbekend", "weet niet" of "wil niet zeggen" opleveren bestaan de twee bestanden respectievelijk uit 1.126 (LVO) en 2.690 waarnemingen (PAP). Het herwegen of ophogen van de steekproeven tot de populatie werd, overeenkomstig de werkwijze van het CBS, gebaseerd op "bouwjaar auto" en "brandstofsoort auto". Jaarkilometrages werden verkregen door maandkilometrages met 12 te vermenigvuldigen.

Transportprobleem

Voor lineaire programmeringsproblemen was standaardprogrammatuur beschikbaar op de IBM 3090 machine van SARA (Stichting Academisch Rekencentrum Amsterdam). Een speciaal hiervoor ontwikkeld pakket is IBM's MPSX/370. Dit pakket is geschikt voor LP en MIP problemen (*Mixed Integer Programming*, waarbij bepaalde variabelen geheeltallig zijn). Het pakket bleek echter minder geschikt voor het oplossen van het onderhavige probleem, omdat het invoeren van de variabelen (1.126 x 2.690 = 3.028.940 stuks) en restricties (1.126 + 2.690 = 3.816 stuks) te bewerkelijk was. Ook het statistische pakket SAS bevat in de module *Operations Research* een algoritme voor transportproblemen, maar daarvoor is ons vraagstuk te groot. De bibliotheek NAG bevat ook een algoritme voor transportproblemen. De subroutine H03ABF voor *Classical Transportation ("Hitchcock") problems* bleek zeer geschikt voor dit geval. Na een aantal aanpassingen - waarbij onder andere (door vectorisering) de iteratiesnelheid werd verdubbeld - werd in iets minder dan vier uur tijd de oplossing gevonden. Exclusief alle voorbereidingen vroeg de oplossing 13.609,78 CPU seconden (= 3 uur en 47 minuten) plus 11.632 vector seconden. Er was 34 Mb geheugenruimte nodig (inclusief 16 Mb voor het besturingssysteem), en de oplossing werd bereikt na 13.449 iteraties.[3)]

Lineair Toewijzingsprobleem

In de literatuur zijn voor Lineaire Toewijzingsproblemen verschillende algoritmen bekend. Wij kozen LAPJV van Jonker en Volgenant (1987). Dit wordt beschouwd als één van de snelste algoritmen. Er was een implementatie beschikbaar voor gebruik op de PC onder Turbo PASCAL. LAPJV had voor de toewijzing van 1.126 van de 2.690 waarnemingen uit het donorbestand aan het recipiëntbestand circa anderhalf uur nodig, bij gebruikmaking van een LASER 286/3, 16 Hz AT personal computer. Wel kwam de omvang van het huidige probleem akelig dicht bij de limiet van LAPJV onder Turbo PASCAL (versie 5.5). Het programma kon niet binnen Turbo zelf draaien, maar moest in gecompileerde vorm buiten Turbo om worden aangeroepen, aangezien Turbo anders zelf te veel "stack-memory" innam.[4)] Wellicht lost het recenter door Volgenant en Warmerdam ontwikkelde vervolg op LAPJV, dat speciaal voor grote LAP's is geschreven, deze problemen op (zie Volgenant, 1991). Dit algoritme LAPMOD levert volgens de auteurs bovendien een aanzienlijke tijdwinst op. Pogingen om ons koppelingsprobleem met dit algoritme op te lossen waren echter nog niet succesvol.

---

3 Voor hulp bij het gebruik van NAG en voor het uitdenken en aanbrengen van de verbeteringen in de subroutine zijn wij veel dank verschuldigd aan drs. J.J.A. Koot en Ir. R. Veldhuyzen van Zanten van SARA.

4 Voor hun hulp bij de toepassing van LAPJV op ons probleem zijn we veel dank verschuldigd aan drs. J.B.J.M. de Kort en dr. A. Volgenant van de faculteit der Economische Wetenschappen en Econometrie van de Universiteit van Amsterdam.

5 Transportprobleem of LAP

In Paragraaf 3 hebben wij al aangegeven dat toepassing van het lineaire toewijzingsprobleem op koppeling doorgaans tot verlies van informatie leidt; bij ongelijke omvang van de bestanden blijft een deel van het donorbestand ongebruikt. Het gevolg voor de marginale verdelingen van de unieke variabelen in het gekoppelde bestand is dat de verdelingen f(x) en f(z) uit D in het gekoppelde bestand niet precies worden gereproduceerd. Koppeling met het transportprobleem heeft dit bezwaar niet: de marginale verdelingen van het gekoppelde bestand zijn dezelfde als in de oorspronkelijke bestanden.

Ter illustratie bevat Tabel 1 de verdeling van het jaarkilometrage in het PAP en in de twee gekoppelde bestanden. Uit deze tabel is op te maken dat waarnemingen uit de klassen met de grootste frequentie in het PAP in het gekoppelde "LAP-bestand" nog eens extra vaak zijn toegewezen aan waarnemingen uit het recipiënt-bestand. De verdeling van de variabele in het "TRANSPORT-bestand" is, zoals verwacht, exact gelijk aan de verdeling in het bestand van herkomst.

Tabel 1 Verdeling van het aantal verreden autokilometers per jaar in het PAP en in het gekoppelde bestand op basis van het transportprobleem (TRANSPORT) en het lineaire toewijzingsprobleem (LAP).

Aantal verreden auto-kilometers per jaar	PAP %	TRANSPORT %	LAP %
≤ 5.000 km	12,3	12,3	10,1
5.000 - 10.000 km	22,4	22,4	25,5
10.000 - 15.000 km	22,8	22,8	24,1
15.000 - 20.000 km	14,0	14,0	15,0
20.000 - 25.000 km	9,3	9,3	9,7
25.000 - 30.000 km	5,8	5,8	5,2
30.000 - 35.000 km	3,6	3,6	2,7
35.000 - 40.000 km	2,5	2,5	2,5
40.000 - 45.000 km	1,6	1,6	1,1
45.000 - 50.000 km	1,1	1,1	0,7
50.000 - 55.000 km	0,5	0,5	0,1
55.000 - 60.000 km	0,5	0,5	0,6
> 60.000 km	1,6	1,6	1,2
Weet niet / geen antwoord	2,0	2,0	1,5
	100,0	100,0	100,0

Wij presenteren in Tabel 2 de resultaten van de twee gekoppelde bestanden voor de gestelde vraag, namelijk hoe hangen opleidingsniveau en kilometrage samen. Bovendien zijn de steekproefaantallen N_s van de gekoppelde bestanden en de - voor beide bestanden gelijke - (opgehoogde) populatie-aantallen N_p vermeld, alle inclusief de (enkele) gevallen waarin het kilometrage onbekend is.

Uit de tabel blijkt dat de uitkomsten in beide bestanden redelijk met elkaar overeenstemmen. De rangorde van opleidingsklassen naar het aantal verreden kilometers is praktisch dezelfde: in beide bestanden neemt het kilometrage toe naarmate de opleiding hoger is. Alleen staan in het "TRANSPORT-bestand" de personen met MBO. MTS en MEAO ten opzichte van de groep HAVO, VWO niet in de juiste volgorde. In het "LAP-bestand" liggen de jaarkilometrages over bijna de hele linie lager dan in het "TRANSPORT-bestand". Dit was eigenlijk al uit Tabel 1 op te maken: het lineaire toewijzingsalgoritme had verhoudingsgewijs vaker auto's met een laag kilometrage uit het PAP toegewezen aan de waarnemingen uit het LVO-bestand.

Tabel 2 Gemiddeld jaarkilometrage naar hoogst genoten opleiding van de Nederlandse automobilist aan de hand van het gekoppelde bestand, geconstrueerd door oplossing van het transportprobleem (TRANSPORT) respectievelijk het lineaire toewijzingsprobleem (LAP).

	TRANSPORT		LAP		
Hoogst genoten opleiding hoofdgebruiker auto	jaarkm auto	N_s	jaarkm auto	N_s	N_p
Lager onderwijs	11.827	207	11.273	64	284.012
LBO/LTS/LEAO e.d.	14.908	894	14.269	264	1.175.856
MAVO/IVO/3 jr VWO e.d.	15.769	553	14.536	164	732.002
MBO/MTS/MEAO e.d.	17.011	752	15.443	223	1.001.040
HAVO/VWO/HBS e.d.	16.061	294	16.149	82	375.670
HBO/WO-kandidaats e.d.	18.172	787	17.212	231	1.074.109
WO-doctoraal e.d.	18.727	325	17.539	97	438.917
Onbekend	31.200	3	36.000	1	5.080
Totaal	16.381	3.815	15.427	1.126	5.086.686

NB: Gegeven een transportprobleem met m "aanvoerpunten" en n "vraagpunten" zijn er m×n mogelijke verbindingen. Bewezen kan worden dat een optimale oplossing van het transportprobleem m+n-1 verbindingen gebruikt (zie bijv Dantzig (1963)).

6 Wel of niet koppelen

Wij concluderen dat met de huidige programmatuur en rekenapparatuur de optimale koppeling van vrij grote databestanden technisch uitvoerbaar is. Nieuwe mainframe-computers moeten in staat worden geacht ook de synthetische koppeling van veel grotere microdata-bestanden tot stand te brengen. Indien een mainframe-computer niet voorhanden is kan op een krachtige PC met algoritmen zoals LAPJV en LAPMOD altijd een "second-best" koppeling worden uitgevoerd, in de vorm van een lineair toewijzingsprobleem.

Hoe is dan toch de terughoudendheid met betrekking tot de praktijk van de synthetische koppeling te verklaren? Behalve de onbekendheid met de materie zullen vooral de problemen in de fase vóór de daadwerkelijke koppeling een reden kunnen zijn. Het transformeren van de beschikbare bestanden, indien al mogelijk, naar een gemeenschappelijke onderzoekseenheid (huishouden, individu, personenauto) en het onderzoek naar de vergelijkbaarheid van de koppelvariabelen (zelfde definities, vergelijkbare vraagstelling in de enquêtes etc.) moet niet te licht worden opgevat. Bovendien moet voor elke nieuwe onderzoeksvraag weer een nieuwe koppeling tot stand worden gebracht - wellicht met andere koppelvariabelen en andere gewichten in de afstandsfunctie (10).

Ten slotte vereisen vele onderzoekingen causale analyses, en hiervoor zijn synthetisch gekoppelde bestanden niet geschikt. De enige oplossing is dan om een geheel nieuwe enquête op te zetten, waarin al de informatie wordt verzameld die men voor het onderzoek nodig denkt te hebben. Dit is erg kostbaar, maar dat zijn enquêtes zoals het LVO of het PAP óók.

Literatuur

Barr, R.S. (1981), Design of experiments to investigate joint distributions in microanalytic simulations, Computer Science and Statistics, Proceedings of the 13th Symposium on the Interface, Springer, p. 17-21.

Barr, R.S. en J.S. Turner (1981), Microdata file merging through large-scale network technology, Mathematical Programming Study 15, p. 1-22.

Bronner, A.E., The end of fusion fear? (1989), paper prepared for AGB-seminar, september 28/29, London; Veldkamp/Marktonderzoek, Amsterdam.

Cramer, J.S. en A.H. Paape (1990), Synthetische koppeling van microdata, deel I: Methoden van koppeling en gebruik van door koppeling verkregen bestanden, deel II: Verkeerskundige bestanden en hun bruikbaarheid voor koppeling; SEO-rapport nr 258, Stichting voor Economisch Onderzoek der Universiteit van Amsterdam (in opdracht van het Projectbureau Integrale Verkeers- en Vervoerstudies), Amsterdam.

Dantzig, G.B. (1963), Linear programming and extensions. Princeton, Princeton University Press.

Jong, W. de (1990), Exact en synthetisch koppelen, of: hoe bedrieglijk overeenkomsten kunnen zijn, intern rapport, Centraal Bureau voor de Statistiek, Voorburg.

Jonker, R. en A. Volgenant (1987), A shortest augmenting path algorithm for dense and sparse linear assignment problems, Computing 38, p. 325-340.

Kadane, J.B., Some statistical problems in merging data files, 1978 Compendium of Tax Research (1978), Office of Tax Analysis, Department of Treasury, Washington, p. 159-179.

Sims, C.A. (1972), Comments, Annals of Economic and Social Measurement, Vol. 1, nr 3, p. 343-346.

Volgenant, A. (1991), On Solving Large Assignment Problems, Instituut voor Actuariaat en Econometrie, Universiteit van Amsterdam.

Ontvangen: 24-6-1992
Geaccepteerd: 18-11-1993